CLAIMS

1. A method of extracting relevant data, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes selected data of the first document, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

determining an edit sequence between at least part of the first set of data and at least part of the second set of data, the edit sequence including any of insertions, deletions, and substitutions; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the edit sequence.

2. The method of claim 1, wherein the edit sequence includes none of insertions, deletions, and substitutions.

3. The method of claim 1, wherein the edit sequence includes at least one of one or more insertions, one or more deletions, and one or more substitutions.

4. The method of claim 1, wherein the edit sequence is at least partly determined by calculating a total cost, and each of one or more of insertions, deletions, substitutions, and matches is associated with one or more costs.

5. The method of claim 4, wherein the one or more costs are at least partly set to encourage the edit sequence to include one or more matches between at least some markup language from the selected data of the first document and at least some markup language from the second document, the markup language including text-based content and tags.

6. The method of claim 4, wherein a first cost is associated with a first match at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second match at a second distance from a root of a tree representation of some set of

data, the first distance is less than the second distance, and the first cost and the second cost are set to encourage the first match more than the second match.

7. The method of claim 4, wherein a first cost is associated with a first insertion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second insertion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

8. The method of claim 4, wherein a first cost is associated with a first deletion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second deletion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

9. The method of claim 4, wherein a first cost is associated with a first substitution at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second substitution at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

10. The method of claim 4, wherein a first cost is associated with a first text-based content substitution such that a first length of substituting text-based content is substantially equal to a first length of substituted text-based content, a second cost is associated with a second text-based content substitution such that a second length of substituting text-based content is substantially different from a second length of substituted text-based content, and the first cost and the second cost are set to discourage the second text-based content substitution more than the first text-based content substitution.

11. The method of claim 4, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of text-based content for one or more tags.

12. The method of claim 4, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of one or more tags for text-based content.

13. The method of claim 4, wherein a first cost is associated with preserving a first tag with unchanged attributes, a second cost is associated with preserving a second tag with one or more changed attributes, and the first cost and the second cost are set to discourage preserving the second tag more than preserving the first tag.

14. The method of claim 1, wherein document data is at least partly from the first document.

15. The method of claim 1, wherein document data is at least partly from the second document.

16. The method of claim 1, wherein the second document is received if the second document is different from the first document.

17. The method of claim 1, wherein the markup language includes at least HTML (Hypertext Markup Language).

18. The method of claim 1, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

19. The method of claim 1 wherein the markup language includes at least WML (Wireless Markup Language).

20. The method of claim 1, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

21. The method of claim 1, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

22. The method of claim 1, wherein the correspondence is at least partly found by one or more of: determining the edit sequence, at least part of at least one of a first plurality of paths from a root of a tree representation of the first set of data to selected data of the tree representation of the first set of data, at least part of at least one of a second plurality of paths from a root of a tree representation of the second set of data to corresponding data of the tree representation of the second set of data, and one or more edit sequences between at least one of the first plurality of paths and at least one of the second plurality of paths.

23. The method of claim 1, wherein one or more of the first set of data and the second set of data is represented at least partly by a tree.

24. The method of claim 1, wherein one or more of the first set of data and the second set of data is represented at least partly by a set of linearized tokens.

25. The method of claim 1, wherein at least the first document and the second document represent different documents.

26. The method of claim 1, wherein the first document and the second document represent a same document.

27. The method of claim 1, wherein the first document and the second document represent different versions of a same document.

28. A method of extracting relevant data, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes selected data of the first document, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

determining a tree-based edit sequence between at least part of the first set of data and at least part of the second set of data, the tree-based edit sequence including any of insertions, deletions, and substitutions; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the tree-based edit sequence.

29. The method of claim 28, wherein the tree-based edit sequence includes none of insertions, deletions, and substitutions.

30. The method of claim 28, wherein the tree-based edit sequence includes at least one of one or more insertions, one or more deletions, and one or more substitutions.

31. The method of claim 28, wherein the tree-based edit sequence is at least partly determined by calculating a total cost, and each of one or more of insertions, deletions, substitutions, and matches is associated with one or more costs.

32. The method of claim 31, wherein the one or more costs are at least partly set to encourage the tree-based edit sequence to include one or more matches between at least some markup language from the selected data of the first document and at least some markup language from the second document, the markup language including text-based content and tags.

33. The method of claim 31, wherein a first cost is associated with a first match at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second match at a second distance from a root of a tree representation of

some set of data, the first distance is less than the second distance, and the first cost and the second cost are set to encourage the first match more than the second match.

34. The method of claim 31, wherein a first cost is associated with a first insertion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second insertion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

35. The method of claim 31, wherein a first cost is associated with a first deletion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second deletion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

36. The method of claim 31, wherein a first cost is associated with a first substitution at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second substitution at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

37. The method of claim 31, wherein a first cost is associated with a first text-based content substitution such that a first length of substituting text-based content is substantially equal to a first length of substituted text-based content, a second cost is associated with a second text-based content substitution such that a second length of substituting text-based content is substantially different from a second length of substituted text-based content, and the first cost and the second cost are set to discourage the second text-based content substitution more than the first text-based content substitution.

38. The method of claim 31, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of text-based content for one or more tags.

39. The method of claim 31, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of one or more tags for text-based content.

40. The method of claim 31, wherein a first cost is associated with preserving a first tag with unchanged attributes, a second cost is associated with preserving a second tag with one or more changed attributes, and the first cost and the second cost are set to discourage preserving the second tag more than preserving the first tag.

41. The method of claim 28, wherein document data is at least partly from the first document.

42. The method of claim 28, wherein document data is at least partly from the second document.

43. The method of claim 28, wherein the second document is received if the second document is different from the first document.

44. The method of claim 28, wherein the markup language includes at least HTML (Hypertext Markup Language).

45. The method of claim 28, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

46. The method of claim 28, wherein the markup language includes at least WML (Wireless Markup Language).

47. The method of claim 28, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

48. The method of claim 28, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

49. The method of claim 28, further comprising:

if two or more corresponding data are found, then:

selecting larger selected data, at least part of the larger selected data including a larger subtree in a first tree representation of the first set of data, the larger subtree including the selected data;

determining a second edit sequence between at least part of the first set of data and at least part of a second tree representation of the second tree of data, the first set of data including at least part of the larger selected data, the second edit sequence including any of insertions, deletions, and substitutions;

finding corresponding data of the second set of data, the corresponding data having a correspondence to the larger selected data, the correspondence at least partly found by determining the second edit sequence; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the second edit sequence.

50. The method of claim 28, wherein the correspondence is at least partly found by one or more of: determining the tree-based edit sequence, at least part of at least one of a first plurality of paths from a root of a tree representation of the first set of data to selected data of the tree representation of the first set of data, at least part of at least one of a second plurality of paths from a root of a tree representation of the second set of data to corresponding data of the tree representation of the second set of data, and one or more tree-based edit sequences between at least one of the first plurality of paths and at least one of the second plurality of paths.

51. The method of claim 28, wherein one or more of the first set of data and the second set of data is represented at least partly by a tree.

52. The method of claim 28, wherein one or more of the first set of data and the second set of data is represented at least partly by a set of linearized tokens.

53. The method of claim 28, wherein at least the first document and the second document represent different documents.

54. The method of claim 28, wherein the first document and the second document represent a same document.

55. The method of claim 28, wherein the first document and the second document represent different versions of a same document.

56. The method of claim 28, further comprising:

determining at least one edit sequence of forward and backward edit sequences between at least part of a first tree representation of the first set of data and at least part of a second tree representation of the second set of data;

performing at least one of 1) and 2):

1a) pruning a relevant subtree from at least part of the first tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

1b) determining a pruned edit sequence between the pruned relevant subtree and at least part of the second tree representation;

2a) pruning a relevant subtree from at least part of the second tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

2b) determining a pruned edit sequence between at least part of the first tree representation and the pruned relevant subtree; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the pruned edit sequence.

57. A method of extraction, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes selected data, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

determining document data of the second set of data, by finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data of the first set of data;

identifying the corresponding data of the second set of data as selected data of the second set of data, the selected data at least partly specifying document data;

accessing at least a third set of data of a third document, the third document including markup language; and

determining document data of the third set of data, by finding corresponding data of the third set of data, the corresponding data having a correspondence to at least one of the selected data of the first set of data and the selected data of the second set of data.

58. The method of claim 57, wherein subsequent sets of data of documents are received, the documents including markup language, document data of the subsequent sets of data are determined by finding corresponding data of the subsequent sets of data, the corresponding data of the subsequent sets correspond to the selected data of earlier sets of data, the corresponding data of the subsequent sets are identified as selected data of the subsequent sets of data, the selected data of the subsequent sets of data at least partly specifying document data, and at least one of selected data of the earlier sets and the selected data of the subsequent data at least partly determine corresponding data of later sets of data, the earlier sets of data are received earlier than the subsequent sets of data, and the later sets of data are received later than the subsequent sets of data.

59. The method of claim 57, wherein document data is at least partly from the first document.

60. The method of claim 57, wherein document data is at least partly from the second document.

61. The method of claim 57, wherein document data is at least partly from the third document.

62. The method of claim 57, wherein the second document is received if the second document is different from the first document.

63. The method of claim 57, wherein the markup language includes at least HTML (Hypertext Markup Language).

64. The method of claim 57, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

65. The method of claim 57, wherein the markup language includes at least WML (Wireless Markup Language).

66. The method of claim 57, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

67. The method of claim 57, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

68. The method of claim 57, wherein at least two of the first document, the second document, and the third document represent different documents.

69. The method of claim 57, wherein at least two of the first document, the second document, and the third document represent a same document.

70. The method of claim 57, wherein at least two of the first document, the second document, and the third document represent different versions of a same document.

71. A method of extraction, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes selected data, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

finding one or more sets of corresponding data of the second set of data, each of one or more sets of corresponding data having a strength of correspondence to the selected data of the first set of data;

if two or more sets of corresponding data are found, then 1) if one of the corresponding sets of data has a substantially higher strength of correspondence than strengths of correspondence of the other corresponding sets of data, assigning a high measure of quality to the selection of the selected data, and 2) assigning a low measure of quality to the selection of the selected data, if at least one of: 2a) none of the corresponding sets of data has a substantially higher strength of correspondence than strengths of correspondence of the other corresponding sets of data, and 2b) if strengths of correspondence of all corresponding sets of data are low.

72. The method of claim 71, wherein document data is at least partly from the first document.

73. The method of claim 71, wherein document data is at least partly from the second document.

74. The method of claim 71, wherein the second document is received if the second document is different from the first document.

75. The method of claim 71, wherein the markup language includes at least HTML (Hypertext Markup Language).

76. The method of claim 71, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

77. The method of claim 71, wherein the markup language includes at least WML (Wireless Markup Language).

78. The method of claim 71, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

79. The method of claim 71, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

80. The method of claim 71, wherein the first document and the second document represent different documents.

81. The method of claim 71, wherein the first document and the second document represent a same document.

82. The method of claim 71, wherein the first document and the second document represent different versions of a same document.

83. A method of extraction, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes a first selected subset and a second selected subset, such that the second selected subset of data is a subset of the first selected subset of data, the first selected subset at least partly specifying document data, the second selected subset at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

determining a first edit sequence between at least part of the first set of data and at least part of the second set of data, the first edit sequence including any of insertions, deletions, and substitutions;

finding a first corresponding subset of the second set of data, the first corresponding subset having a correspondence to the first selected subset, the correspondence at least partly found by determining the first edit sequence;

determining a second edit sequence between at least part of the first set of data and at least part of the second set of data, the first set of data including at least part of the first selected subset, the second set of data including at least part of the first corresponding subset, the second edit sequence including any of insertions, deletions, and substitutions; and

finding a second corresponding subset of the second set of data, the second corresponding subset having a correspondence to the second selected subset, the correspondence at least partly found by determining the second edit sequence.

84. The method of claim 83, wherein at least one of the first edit sequence and the second edit sequence includes none of insertions, deletions, and substitutions.

85. The method of claim 83, wherein at least one of the first edit sequence and the second edit sequence includes at least one of one or more insertions, one or more deletions, and one or more substitutions.

86. The method of claim 83, wherein at least one of the first edit sequence and the second edit sequence is at least partly determined by calculating a total cost, and each of one or more of insertions, deletions, substitutions, and matches is associated with one or more costs.

87. The method of claim 86, wherein the one or more costs are at least partly set to encourage the edit sequence to include one or more matches between at least some markup language from the selected data of the first document and at least some markup language from the second document, the markup language including text-based content and tags.

88. The method of claim 86, wherein a first cost is associated with a first match at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second match at a second distance from a root of a tree representation of

some set of data, the first distance is less than the second distance, and the first cost and the second cost are set to encourage the first match more than the second match.

89. The method of claim 86, wherein a first cost is associated with a first insertion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second insertion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

90. The method of claim 86, wherein a first cost is associated with a first deletion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second deletion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

91. The method of claim 86, wherein a first cost is associated with a first substitution at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second substitution at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

92. The method of claim 86, wherein a first cost is associated with a first text-based content substitution such that a first length of substituting text-based content is substantially equal to a first length of substituted text-based content, a second cost is associated with a second text-based content substitution such that a second length of substituting text-based content is substantially different from a second length of substituted text-based content, and the first cost and the second cost are set to discourage the second text-based content substitution more than the first text-based content substitution.

93. The method of claim 86, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of text-based content for one or more tags.

94. The method of claim 86, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of one or more tags for text-based content.

95. The method of claim 86, wherein a first cost is associated with preserving a first tag with unchanged attributes, a second cost is associated with preserving a second tag with one or more changed attributes, and the first cost and the second cost are set to discourage preserving the second tag more than preserving the first tag.

96. The method of claim 83, wherein document data is at least partly from the first document.

97. The method of claim 83, wherein document data is at least partly from the second document.

98. The method of claim 83, wherein the second document is received if the second document is different from the first document.

99. The method of claim 83, wherein the markup language includes at least HTML (Hypertext Markup Language).

100. The method of claim 83, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

101. The method of claim 83, wherein the markup language includes at least WML (Wireless Markup Language).

102. The method of claim 83, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

103. The method of claim 83, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

104. The method of claim 83, further comprising:

if two or more corresponding data are found, then:

selecting larger selected data, at least part of the larger selected data including a larger subtree in a first tree representation of the first set of data, the larger subtree including the selected data;

determining a third edit sequence between at least part of the first set of data and at least part of a second tree representation of the second set of data, the first set of data including at least part of the larger selected data, the third edit sequence including any of insertions, deletions, and substitutions;

finding corresponding data of the second set of data, the corresponding data having a correspondence to the larger selected data, the correspondence at least partly found by determining the third edit sequence; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the third edit sequence.

105. The method of claim 83, wherein one or more of the first set of data and the second set of data is represented at least partly by a tree.

106. The method of claim 83, wherein one or more of the first set of data and the second set of data is represented at least partly by a set of linearized tokens.

107. The method of claim 83, wherein the first document and the second document represent different documents.

108. The method of claim 83, wherein the first document and the second document represent a same document.

109. The method of claim 83, wherein the first document and the second document represent different versions of a same document.

110. The method of claim 83, wherein at least one of the first edit sequence and the second edit sequence includes a tree-based edit sequence.

111. The method of claim 83, wherein at least one of determining the first edit sequence and determining the second edit sequence comprises:

determining at least one edit sequence of forward and backward edit sequences between at least part of a first tree representation of the first set of data and at least part of a second tree representation of the second set of data;

performing at least one of 1) and 2):

1a) pruning a relevant subtree from at least part of the first tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

1b) determining a pruned edit sequence between the pruned relevant subtree and at least part of the second tree representation;

2a) pruning a relevant subtree from at least part of the second tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

2b) determining a pruned edit sequence between at least part of the first tree representation and the pruned relevant subtree; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the pruned edit sequence.

112. A method of extraction, comprising:

accessing at least a plurality of first sets of data of a plurality of first documents, the first documents including markup language, wherein each of the plurality of first sets of data includes selected data, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

determining a most corresponding first set of data of the plurality of first sets of data, the most corresponding first set of data having most correspondence with the second set of data, by comparing partial representations of the plurality of first sets of data with a partial representation of the second set of data.

113. The method of claim 112, wherein document data is at least partly from one or more of the plurality of first documents.

114. The method of claim 112, wherein document data is at least partly from the second document.

115. The method of claim 112, wherein the second document is received if the second document is different from at least one of the plurality of first documents.

116. The method of claim 112, wherein the second document is received if the second document is different from all of the plurality of first documents.

117. The method of claim 112, wherein the markup language includes at least HTML (Hypertext Markup Language).

118. The method of claim 112, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

119. The method of claim 112, wherein the markup language includes at least WML (Wireless Markup Language).

120. The method of claim 112, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

121. The method of claim 112, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

122. The method of claim 112, wherein the partial representation of the second set of data includes a hash value computed on at least part of the second set of data.

123. The method of claim 112, wherein a partial representation of a first set of data of the plurality of first sets of data includes a hash value computed on at least part of the first set of data of the plurality of first sets of data.

124. The method of claim 112, wherein the partial representation of the second set of data includes at least a partial syntax tree of the second set of data.

125. The method of claim 112, wherein a partial representation of a first set of data of the plurality of first sets of data includes at least a partial syntax tree of the first set of data of the plurality of first sets of data.

126. The method of claim 112, wherein the partial representation of the second set of data includes a hash value computed on at least a partial syntax tree of the second set of data.

127. The method of claim 112, wherein a partial representation of a first set of data of the plurality of first sets of data includes a hash value computed on at least a partial syntax tree of the first set of data of the plurality of first sets of data.

128. The method of claim 112, wherein the partial representation of the second set of data includes at least one of a part of a name of the second set of data and a part of a name of the second document.

129. The method of claim 112, wherein a partial representation of a first set of data of the plurality of first sets of data of first documents includes at least one of 1) a part of a

name of the first set of data of the plurality of first sets of data and 2) a part of a name of a first document of the first documents, the first document of the first documents including the first set of data of the plurality of first sets of data.

130. The method of claim 112, wherein at least two documents out of the first plurality of documents and the second document represent different documents.

131. The method of claim 112, wherein at least two documents out of the first plurality of documents and the second document represent a same document.

132. The method of claim 112, wherein at least two documents out of the first plurality of documents and the second document represent different versions of a same document.

133. A method of extraction, comprising:

accessing at least a first tree of data of a first document, the first document including markup language, wherein the first tree of data includes selected data, the selected data at least partly specifying document data;

accessing at least a second tree of data of a second document, the second document including markup language;

determining at least one edit sequence of forward and backward edit sequences between at least part of the first tree and at least part of the second tree;

performing at least one of 1) and 2):

1a) pruning a relevant subtree from at least part of the first tree, the relevant subtree at least partly determined from the forward and backward edit sequences;

1b) determining a pruned edit sequence between the pruned relevant subtree and at least part of the second tree;

2a) pruning a relevant subtree from at least part of the second tree, the relevant subtree at least partly determined from the forward and backward edit sequences;

2b) determining a pruned edit sequence between at least part of the first tree and the pruned relevant subtree; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the pruned edit sequence.

134. The method of claim 133, wherein document data is at least partly from the first document.

135. The method of claim 133, wherein document data is at least partly from the second document.

136. The method of claim 133, wherein the second document is received if the second document is different from the first document.

137. The method of claim 133, wherein the markup language includes at least HTML (Hypertext Markup Language).

138. The method of claim 133, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

139. 11000, wherein the markup language includes at least WML (Wireless Markup Language).

140. The method of claim 133, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

141. The method of claim 133, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

142. The method of claim 133, wherein determining forward and backward edit sequences, pruning a relevant subtree, and determining a pruned edit sequence are

performed for each of a plurality of subtree pairs, each of the plurality of subtree pairs including a subtree from the first tree and a subtree from the second tree.

143. The method of claim 133, wherein the first document and the second document represent different documents.

144. The method of claim 133, wherein the first document and the second document represent a same document.

145. 133, wherein the first document and the second document represent different versions of a same document.

146. A method of extracting relevant data, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes selected data of the first document, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

determining a first edit sequence between at least part of the first set of data and at least part of the second set of data, the first edit sequence including any of insertions, deletions, and substitutions;

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the first edit sequence;

if two or more corresponding data are found, then:

selecting larger selected data, at least part of the larger selected data including a larger subtree in a tree representation of the first set of data, the larger subtree including the selected data;

determining a second edit sequence between at least part of the first set of data and at least part of the second set of data, the first set of data including at least part of the larger selected data, the second edit sequence including any of insertions, deletions, and substitutions;

finding corresponding data of the second set of data, the corresponding data having a correspondence to the larger selected data, the correspondence at least partly found by determining the second edit sequence; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the second edit sequence.

147. The method of claim 146, wherein document data is at least partly from the first document.

148. The method of claim 146, wherein document data is at least partly from the second document.

149. The method of claim 146, wherein the second document is received if the second document is different from the first document.

150. The method of claim 146, wherein the markup language includes at least HTML (Hypertext Markup Language).

151. The method of claim 146, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

152. The method of claim 146, wherein the markup language includes at least WML (Wireless Markup Language).

153. The method of claim 146, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

154. The method of claim 146, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

155. The method of claim 146, wherein the first document and the second document represent different documents.

156. The method of claim 146, wherein the first document and the second document represent a same document.

157. The method of claim 146, wherein the first document and the second document represent different versions of a same document.

158. A method of extraction, comprising:

accessing at least a first tree of data of a first document, the first document including markup language, wherein the first tree of data includes selected data, the selected data at least partly specifying document data;

accessing at least a second tree of data of a second document, the second document including markup language;

performing tree traversal on at least part of the second tree, the tree traversal at least partly guided by the selected data and by at least part of the first tree; and

if tree traversal fails due to one or more differences between at least part of the second tree and at least part of the selected data, then:

determining an edit sequence between at least part of the second tree and at least part of the first tree, the first tree including at least part of the selected data;

finding corresponding data for at least part of the second tree, the corresponding data having a correspondence to at least part of the selected data, the correspondence at least partly found by determining the edit sequence; and

continuing to perform tree traversal on at least part of the second tree, the tree traversal at least partly guided by the corresponding data.

159. The method of claim 158, wherein for subsequent set tree traversal failures, determining, finding and continuing are repeated.

160. The method of claim 158, wherein document data is at least partly from the first document.

161. The method of claim 158, wherein document data is at least partly from the second document.

162. The method of claim 158, wherein the second document is received if the second document is different from the first document.

163. The method of claim 158, wherein the markup language includes at least HTML (Hypertext Markup Language).

164. The method of claim 158, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

165. The method of claim 158, wherein the markup language includes at least WML (Wireless Markup Language).

166. The method of claim 158, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

167. The method of claim 158, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

168. The method of claim 158, wherein the first document and the second document represent different documents.

169. The method of claim 158, wherein the first document and the second document represent a same document.

170. The method of claim 158, wherein the first document and the second document represent different versions of a same document.

171. A method of extracting relevant data, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes selected data of the first document, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

determining an edit sequence between the first set of data and the second set of data, the edit sequence including any of insertions, deletions, and substitutions; and

if the edit sequence fails a test, determining a tree-based edit sequence between the first set of data and the second set of data, the edit sequence including any of insertions, deletions, and substitutions.

172. The method of claim 171, wherein document data is at least partly from the first document.

173. The method of claim 171, wherein document data is at least partly from the second document.

174. The method of claim 171, wherein the second document is received if the second document is different from the first document.

175. The method of claim 171, wherein the markup language includes at least HTML (Hypertext Markup Language).

176. The method of claim 171, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

177. The method of claim 171, wherein the markup language includes at least WML (Wireless Markup Language).

178. The method of claim 171, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

179. The method of claim 171, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

180. The method of claim 171, wherein the first document and the second document represent different documents.

181. The method of claim 171, wherein the first document and the second document represent a same document.

182. The method of claim 171, wherein the first document and the second document represent different versions of a same document.

183. A method of extraction, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the fi5rst set of data includes selected data, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

determining document data of the second set of data, by finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data of the first set of data, the correspondence at least partly determined by a first edit sequence between at least part of the first set of data and at least part of the second set of data, the first edit sequence including any of insertions, deletions, and substitutions;

identifying the corresponding data of the second set of data as selected data of the second set of data, the selected data at least partly specifying document data;

accessing at least a third set of data of a third document, the third document including markup language; and

determining document data of the third set of data, by finding corresponding data of the third set of data, the corresponding data having a correspondence to at least one of the selected data of the first set of data and the selected data of the second set of data, the correspondence at least partly determined by a second edit sequence between at least part of the third set of data and at least one of at least part of the first set of data and at least part of the second set of data, the second edit sequence including any of insertions, deletions, and substitutions.

184. The method of claim 183, wherein at least one of the first edit sequence and the second edit sequence includes none of insertions, deletions, and substitutions.

185. The method of claim 183, wherein at least one of the first edit sequence and the second edit sequence includes at least one of one or more insertions, one or more deletions, and one or more substitutions.

186. The method of claim 183, wherein at least one of the first edit sequence and the second edit sequence is at least partly determined by calculating a total cost, and each of one or more of insertions, deletions, substitutions, and matches is associated with one or more costs.

187. The method of claim 185, wherein the one or more costs are at least partly set to encourage the edit sequence to include one or more matches between at least some markup language from the selected data of the first document and at least some markup language from the second document, the markup language including text-based content and tags.

188. The method of claim 185, wherein a first cost is associated with a first match at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second match at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are set to encourage the first match more than the second match.

189. The method of claim 185, wherein a first cost is associated with a first insertion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second insertion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

190. The method of claim 185, wherein a first cost is associated with a first deletion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second deletion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

191. The method of claim 185, wherein a first cost is associated with a first substitution at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second substitution at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

192. The method of claim 185, wherein a first cost is associated with a first text-based content substitution such that a first length of substituting text-based content is substantially equal to a first length of substituted text-based content, a second cost is associated with a second text-based content substitution such that a second length of substituting text-based content is substantially different from a second length of substituted text-based content, and the first cost and the second cost are set to discourage the second text-based content substitution more than the first text-based content substitution.

193. The method of claim 185, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of text-based content for one or more tags.

194. The method of claim 185, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of one or more tags for text-based content.

195. The method of claim 185, wherein a first cost is associated with preserving a first tag with unchanged attributes, a second cost is associated with preserving a second tag with one or more changed attributes, and the first cost and the second cost are set to discourage preserving the second tag more than preserving the first tag.

196. The method of claim 183, wherein subsequent sets of data of documents are received, the documents including markup language, document data of the subsequent sets of data are determined by finding corresponding data of the subsequent sets of data, the corresponding data of the subsequent sets correspond to the selected data of earlier sets of data, the corresponding data of the subsequent sets are identified as selected data of the subsequent sets of data, the selected data of the subsequent sets of data at least partly specifying document data, and at least one of selected data of the earlier sets and the selected data of the subsequent data at least partly determine corresponding data of later sets of data, the earlier sets of data are received earlier than the subsequent sets of data, and the later sets of data are received later than the subsequent sets of data.

197. The method of claim 183, wherein document data is at least partly from the first document.

198. The method of claim 183, wherein document data is at least partly from the second document.

199. The method of claim 183, wherein document data is at least partly from the third document.

200. The method of claim 183, wherein the second document is received if the second document is different from the first document.

201. The method of claim 183, wherein the markup language includes at least HTML (Hypertext Markup Language).

202. The method of claim 183, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

203. The method of claim 183, wherein the markup language includes at least WML (Wireless Markup Language).

204. The method of claim 183, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

205. The method of claim 183, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

206. The method of claim 183, further comprising:

if two or more corresponding data are found, then:

selecting larger selected data, at least part of the larger selected data including a larger subtree in a first tree representation of the first set of data, the larger subtree including the selected data;

determining a third edit sequence between at least part of the first set of data and at least part of a second tree representation of the second set of data, the first set of data including at least part of the larger selected data, the third edit sequence including any of insertions, deletions, and substitutions;

finding corresponding data of the second set of data, the corresponding data having a correspondence to the larger selected data, the correspondence at least partly found by determining the third edit sequence; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the third edit sequence.

207. The method of claim 183, wherein one or more of the first set of data and the second set of data is represented at least partly by a tree.

208. The method of claim 183, wherein one or more of the first set of data and the second set of data is represented at least partly by a set of linearized tokens.

209. The method of claim 183, wherein at least two of the first document, the second document, and the third document represent different documents.

210. The method of claim 183, wherein at least two of the first document, the second document, and the third document represent a same document.

211. The method of claim 183, wherein at least two of the first document, the second document, and the third document represent different versions of a same document.

212. The method of claim 183, wherein at least one of the first edit sequence and the second edit sequence includes a tree-based edit sequence.

213. The method of claim 183, wherein determining the edit sequence comprises:

determining at least one edit sequence of forward and backward edit sequences between at least part of a first tree representation of the first set of data and at least part of a second tree representation of the second set of data;

performing at least one of 1) and 2):

1a) pruning a relevant subtree from at least part of the first tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

1b) determining a pruned edit sequence between the pruned relevant subtree and at least part of the second tree representation;

2a) pruning a relevant subtree from at least part of the second tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

2b) determining a pruned edit sequence between at least part of the first tree representation and the pruned relevant subtree; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the pruned edit sequence.

214. A method of extraction, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes selected data, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

finding one or more sets of corresponding data of the second set of data, each of one or more sets of corresponding data having a strength of correspondence to the selected data of the first set of data, the strength of correspondence at least partly determined by an edit sequence between at least part of the second set of data and at least part of the first set of data, the edit sequence including any of insertions, deletions, and substitutions;

if two or more sets of corresponding data are found, then 1) if one of the corresponding sets of data has a substantially higher strength of correspondence than strengths of correspondence of the other corresponding sets of data, assigning a high measure of quality to the selection of the selected data, and 2) if none of the corresponding sets of data has a substantially higher strength of correspondence than strengths of correspondence of the other corresponding sets of data, assigning a low measure of quality to the selection of the selected data.

215. The method of claim 214, wherein the edit sequence includes none of insertions, deletions, and substitutions.

216. The method of claim 214, wherein the edit sequence includes at least one of one or more insertions, one or more deletions, and one or more substitutions.

217. The method of claim 214, wherein the edit sequence is at least partly determined by calculating a total cost, and each of one or more of insertions, deletions, substitutions, and matches is associated with one or more costs.

218. The method of claim 217, wherein the one or more costs are at least partly set to encourage the edit sequence to include one or more matches between at least some markup language from the selected data of the first document and at least some markup language from the second document, the markup language including text-based content and tags.

219. The method of claim 217, wherein a first cost is associated with a first match at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second match at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are set to encourage the first match more than the second match.

220. The method of claim 217, wherein a first cost is associated with a first insertion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second insertion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

221 The method of claim 217, wherein a first cost is associated with a first deletion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second deletion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

222. The method of claim 217, wherein a first cost is associated with a first substitution at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second substitution at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

223. The method of claim 217, wherein a first cost is associated with a first text-based content substitution such that a first length of substituting text-based content is substantially equal to a first length of substituted text-based content, a second cost is associated with a second text-based content substitution such that a second length of substituting text-based content is substantially different from a second length of substituted text-based content, and the first cost and the second cost are set to discourage the second text-based content substitution more than the first text-based content substitution.

224. The method of claim 217, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of text-based content for one or more tags.

225. The method of claim 217, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of one or more tags for text-based content.

226. The method of claim 217, wherein a first cost is associated with preserving a first tag with unchanged attributes, a second cost is associated with preserving a second tag with one or more changed attributes, and the first cost and the second cost are set to discourage preserving the second tag more than preserving the first tag.

227. The method of claim 214, wherein document data is at least partly from the first document.

228. The method of claim 214, wherein document data is at least partly from the second document.

229. The method of claim 214, wherein the second document is received if the second document is different from the first document.

230. The method of claim 214, wherein the markup language includes at least HTML (Hypertext Markup Language).

231. The method of claim 214, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

232. The method of claim 214, wherein the markup language includes at least WML (Wireless Markup Language).

233. The method of claim 214, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

234. The method of claim 214, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

235. The method of claim 214, further comprising:

if two or more corresponding data are found, then:

selecting larger selected data, at least part of the larger selected data including a larger subtree in a first tree representation of the first set of data, the larger subtree including the selected data;

determining a second edit sequence between at least part of the first set of data and at least part of a second tree representation of the second set of data, the first set of data including at least part of the larger selected data, the second edit sequence including any of insertions, deletions, and substitutions;

finding corresponding data of the second set of data, the corresponding data having a correspondence to the larger selected data, the correspondence at least partly found by determining the second edit sequence; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the second edit sequence.

236. The method of claim 214, wherein one or more of the first set of data and the second set of data is represented at least partly by a tree.

237. The method of claim 214, wherein one or more of the first set of data and the second set of data is represented at least partly by a set of linearized tokens.

238. The method of claim 214, wherein the first document and the second document represent different documents.

239. The method of claim 214, wherein the first document and the second document represent a same document.

240. The method of claim 214, wherein the first document and the second document represent different versions of a same document.

241. The method of claim 214, wherein at least one of the first edit sequence and the second edit sequence includes a tree-based edit sequence.

242. The method of claim 214, wherein determining the edit sequence comprises:

determining at least one edit sequence of forward and backward edit sequences between at least part of a first tree representation of the first set of data and at least part of a second tree representation of the second set of data;

performing at least one of 1) and 2):

1a) pruning a relevant subtree from at least part of the first tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

1b) determining a pruned edit sequence between the pruned relevant subtree and at least part of the second tree representation;

2a) pruning a relevant subtree from at least part of the second tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

2b) determining a pruned edit sequence between at least part of the first tree representation and the pruned relevant subtree; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the pruned edit sequence.

243. A method of extraction, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes a first selected subset and a second selected subset, such that the second selected subset of data is a subset of the first selected subset of data, the first selected subset at least partly specifying document data, the second selected subset at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

finding a first corresponding subset of the second set of data, the first corresponding subset having a correspondence to the first selected subset; and

finding a second corresponding subset of the second set of data, the second corresponding subset having a correspondence to the second selected subset.

244. The method of claim 243, wherein document data is at least partly from the first document.

245. The method of claim 243, wherein document data is at least partly from the second document.

246. The method of claim 243, wherein the second document is received if the second document is different from the first document.

247. The method of claim 243, wherein the markup language includes at least HTML (Hypertext Markup Language).

248. The method of claim 243, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

249. The method of claim 243, wherein the markup language includes at least WML (Wireless Markup Language).

250. The method of claim 243, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

251. The method of claim 243, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

252. The method of claim 243, further comprising:

if two or more corresponding data are found, then:

selecting larger selected data, at least part of the larger selected data including a larger subtree in a first tree representation of the first set of data, the larger subtree including the selected data;

determining a first edit sequence between at least part of the first set of data and at least part of a second tree representation of the second set of data, the first set of data including at least part of the larger selected data, the first edit sequence including any of insertions, deletions, and substitutions;

finding corresponding data of the second set of data, the corresponding data having a correspondence to the larger selected data, the correspondence at least partly found by determining the first edit sequence; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the first edit sequence.

253. The method of claim 243, wherein one or more of the first set of data and the second set of data is represented at least partly by a tree.

254. The method of claim 243, wherein one or more of the first set of data and the second set of data is represented at least partly by a set of linearized tokens.

255. The method of claim 243, wherein the first document and the second document represent different documents.

256. The method of claim 243, wherein the first document and the second document represent a same document.

257. The method of claim 243, wherein the first document and the second document represent different versions of a same document.

258. A method of extraction, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes selected data, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

determining document data of the second set of data, by finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data of the first set of data, the correspondence at least partly determined by a first tree-based edit sequence between at least part of the first set of data and at least part of the second set of data, the first tree-based edit sequence including any of insertions, deletions, and substitutions;

identifying the corresponding data of the second set of data as selected data of the second set of data, the selected data at least partly specifying document data;

accessing at least a third set of data of a third document, the third document including markup language; and

determining document data of the third set of data, by finding corresponding data of the third set of data, the corresponding data having a correspondence to at least one of the selected data of the first set of data and the selected data of the second set of data, the

correspondence at least partly determined by a second tree-based edit sequence between at least part of the third set of data and at least one of at least part of the first set of data and at least part of the second set of data, the second tree-based edit sequence including any of insertions, deletions, and substitutions.

259. The method of claim 258, wherein at least one of the first tree-based edit sequence and the second tree-based edit sequence includes none of insertions, deletions, and substitutions.

260. The method of claim 258, wherein at least one of the first tree-based edit sequence and the second tree-based edit sequence includes at least one of one or more insertions, one or more deletions, and one or more substitutions.

261. The method of claim 258, wherein at least one of the first tree-based edit sequence and the second tree-based edit sequence is at least partly determined by calculating a total cost, and each of one or more of insertions, deletions, substitutions, and matches is associated with one or more costs.

262. The method of claim 261, wherein the one or more costs are at least partly set to encourage the tree-based edit sequence to include one or more matches between at least some markup language from the selected data of the first document and at least some markup language from the second document, the markup language including text-based content and tags.

263. The method of claim 261, wherein a first cost is associated with a first match at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second match at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are set to encourage the first match more than the second match.

264. The method of claim 261, wherein a first cost is associated with a first insertion at a first distance from a root of a tree representation of some set of data, a second cost is

associated with a second insertion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

265. The method of claim 261, wherein a first cost is associated with a first deletion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second deletion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

266. The method of claim 261, wherein a first cost is associated with a first substitution at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second substitution at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

267. The method of claim 261, wherein a first cost is associated with a first text-based content substitution such that a first length of substituting text-based content is substantially equal to a first length of substituted text-based content, a second cost is associated with a second text-based content substitution such that a second length of substituting text-based content is substantially different from a second length of substituted text-based content, and the first cost and the second cost are set to discourage the second text-based content substitution more than the first text-based content substitution.

268. The method of claim 261, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of text-based content for one or more tags.

269. The method of claim 261, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of one or more tags for text-based content.

270. The method of claim 261, wherein a first cost is associated with preserving a first tag with unchanged attributes, a second cost is associated with preserving a second tag with one or more changed attributes, and the first cost and the second cost are set to discourage preserving the second tag more than preserving the first tag.

271. The method of claim 258, wherein subsequent sets of data of documents are received, the documents including markup language, document data of the subsequent sets of data are determined by finding corresponding data of the subsequent sets of data, the corresponding data of the subsequent sets correspond to the selected data of earlier sets of data, the corresponding data of the subsequent sets are identified as selected data of the subsequent sets of data, the selected data of the subsequent sets of data at least partly specifying document data, and at least one of selected data of the earlier sets and the selected data of the subsequent data at least partly determine corresponding data of later sets of data, the earlier sets of data are received earlier than the subsequent sets of data, and the later sets of data are received later than the subsequent sets of data.

272. The method of claim 258, wherein document data is at least partly from the first document.

273. The method of claim 258, wherein document data is at least partly from the second document.

274. The method of claim 258, wherein document data is at least partly from the third document.

275. The method of claim 258, wherein the second document is received if the second document is different from the first document.

276. The method of claim 258, wherein the markup language includes at least HTML (Hypertext Markup Language).

277. The method of claim 258, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

278. The method of claim 258, wherein the markup language includes at least WML (Wireless Markup Language).

279. The method of claim 258, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

280. The method of claim 258, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

281. The method of claim 258, further comprising:

if two or more corresponding data are found, then:

selecting larger selected data, at least part of the larger selected data including a larger subtree in a first tree representation of the first set of data, the larger subtree including the selected data;

determining a third tree-based edit sequence between at least part of the first set of data and at least part of a second tree representation of the second set of data, the first set of data including at least part of the larger selected data, the third tree-based edit sequence including any of insertions, deletions, and substitutions;

finding corresponding data of the second set of data, the corresponding data having a correspondence to the larger selected data, the correspondence at least partly found by determining the third tree-based edit sequence; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the third tree-based edit sequence.

282/ The method of claim 258, wherein one or more of the first set of data and the second set of data is represented at least partly by a tree.

283 The method of claim 258, wherein one or more of the first set of data and the second set of data is represented at least partly by a set of linearized tokens.

284. The method of claim 258, wherein at least two of the first document, the second document, and the third document represent different documents.

285. The method of claim 258, wherein at least two of the first document, the second document, and the third document represent a same document.

286. The method of claim 258, wherein at least two of the first document, the second document, and the third document represent different versions of a same document.

287 The method of claim 258, wherein at least one of the first tree-based edit sequence and the second tree-based edit sequence includes a tree-based tree-based edit sequence.

288. The method of claim 258, wherein determining the tree-based edit sequence comprises:

determining at least one tree-based edit sequence of forward and backward edit sequences between at least part of a first tree representation of the first set of data and at least part of a second tree representation of the second set of data;

performing at least one of 1) and 2):

1a) pruning a relevant subtree from at least part of the first tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

1b) determining a pruned tree-based edit sequence between the pruned relevant subtree and at least part of the second tree representation;

2a) pruning a relevant subtree from at least part of the second tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

2b) determining a pruned tree-based edit sequence between at least part of the first tree representation and the pruned relevant subtree; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the pruned tree-based edit sequence.

289. A method of extraction, comprising:

accessing at least a first set of data of a first document, the first document including markup language, wherein the first set of data includes selected data, the selected data at least partly specifying document data;

accessing at least a second set of data of a second document, the second document including markup language;

finding one or more sets of corresponding data of the second set of data, each of one or more sets of corresponding data having a strength of correspondence to the selected data of the first set of data, the strength of correspondence at least partly determined by some tree-based edit sequence between at least part of the second set of data and at least part of the first set of data, the tree-based edit sequence including any of insertions, deletions, and substitutions;

if two or more sets of corresponding data are found, then 1) if one of the corresponding sets of data has a substantially higher strength of correspondence than strengths of correspondence of the other corresponding sets of data, assigning a high measure of quality to the selection of the selected data, and 2) if none of the corresponding sets of data has a substantially higher strength of correspondence than strengths of correspondence of the other corresponding sets of data, assigning a low measure of quality to the selection of the selected data.

290. The method of claim 289, wherein the tree-based edit sequence includes none of insertions, deletions, and substitutions.

291. The method of claim 289, wherein the tree-based edit sequence includes at least one of one or more insertions, one or more deletions, and one or more substitutions.

292. The method of claim 289, wherein the tree-based edit sequence is at least partly determined by calculating a total cost, and each of one or more of insertions, deletions, substitutions, and matches is associated with one or more costs.

293. The method of claim 292, wherein the one or more costs are at least partly set to encourage the tree-based edit sequence to include one or more matches between at least some markup language from the selected data of the first document and at least some markup language from the second document, the markup language including text-based content and tags.

294. The method of claim 292, wherein a first cost is associated with a first match at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second match at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are set to encourage the first match more than the second match.

295. The method of claim 292, wherein a first cost is associated with a first insertion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second insertion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

296. The method of claim 292, wherein a first cost is associated with a first deletion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second deletion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

297. The method of claim 292, wherein a first cost is associated with a first substitution at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second substitution at a second distance from a root of a tree

representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

298. The method of claim 292, wherein a first cost is associated with a first text-based content substitution such that a first length of substituting text-based content is substantially equal to a first length of substituted text-based content, a second cost is associated with a second text-based content substitution such that a second length of substituting text-based content is substantially different from a second length of substituted text-based content, and the first cost and the second cost are set to discourage the second text-based content substitution more than the first text-based content substitution.

299. The method of claim 292, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of text-based content for one or more tags.

300. The method of claim 292, wherein markup language includes at least text-based content and tags, and the one or more costs are at least partly set to discourage substitutions of one or more tags for text-based content.

301. The method of claim 292, wherein a first cost is associated with preserving a first tag with unchanged attributes, a second cost is associated with preserving a second tag with one or more changed attributes, and the first cost and the second cost are set to discourage preserving the second tag more than preserving the first tag.

302. The method of claim 289, wherein document data is at least partly from the first document.

303. The method of claim 289, wherein document data is at least partly from the second document.

304. The method of claim 289, wherein the second document is received if the second document is different from the first document.

305. The method of claim 289, wherein the markup language includes at least HTML (Hypertext Markup Language).

306. The method of claim 289, wherein the markup language includes at least one of XML, a subset of XML, and a specialization of XML (eXtensible Markup Language).

307. The method of claim 289, wherein the markup language includes at least WML (Wireless Markup Language).

308. The method of claim 289, wherein the markup language includes at least one of SGML, a subset of SGML, and a specialization of SGML (Standard Generalized Markup Language).

309. The method of claim 289, wherein the markup language includes at least text-based content and tags, the tags detailing one or more of structure of content, semantics of content, and formatting information about text-based content.

310. The method of claim 289, further comprising:

if two or more corresponding data are found, then:

selecting larger selected data, at least part of the larger selected data including a larger subtree in a first tree representation of the first set of data, the larger subtree including the selected data;

determining a second tree-based edit sequence between at least part of the first set of data and at least part of a second tree representation of the second set of data, the first set of data including at least part of the larger selected data, the second tree-based edit sequence including any of insertions, deletions, and substitutions;

finding corresponding data of the second set of data, the corresponding data having a correspondence to the larger selected data, the correspondence at least partly found by determining the second tree-based edit sequence; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the second tree-based edit sequence.

311. The method of claim 289, wherein one or more of the first set of data and the second set of data is represented at least partly by a tree.

312. The method of claim 289, wherein one or more of the first set of data and the second set of data is represented at least partly by a set of linearized tokens.

313. The method of claim 289, wherein the first document and the second document represent different documents.

314. The method of claim 289, wherein the first document and the second document represent a same document.

315. The method of claim 289, wherein the first document and the second document represent different versions of a same document.

316. The method of claim 289, wherein at least one of the first tree-based edit sequence and the second tree-based edit sequence includes a tree-based tree-based edit sequence.

317. The method of claim 289, wherein determining the tree-based edit sequence comprises:

determining at least one tree-based edit sequence of forward and backward edit sequences between at least part of a first tree representation of the first set of data and at least part of a second tree representation of the second set of data;

performing at least one of 1) and 2):

1a) pruning a relevant subtree from at least part of the first tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

1b) determining a pruned tree-based edit sequence between the pruned relevant subtree and at least part of the second tree representation;

2a) pruning a relevant subtree from at least part of the second tree representation, the relevant subtree at least partly determined from the forward and backward edit sequences;

2b) determining a pruned tree-based edit sequence between at least part of the first tree representation and the pruned relevant subtree; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the pruned tree-based edit sequence.